A Review of Question Answering System in Online Health Guide in Natural Language

Premchand, Yogesh Rai

Abstract— **Generation of answer in QA system isone of the most challenging research areas in natural language and performing in the medical domain is more difficult. The main reason is: patient has faith in doctor's information but may doubt on the machine. Also, accuracy of the system is restricted in natural language. The computer may find the information of a disease but not as efficient as human. In this paper, various well known question answering systems are discussed and analyzed with their benefits and limitations. Finally, in this review, a new approach is visualized to overcome the limitations of present question answering system.**

Keywords— **Medical domain, Natural language, QA**

I. INTRODUCTION

Question Answering (QA)is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. Question Answering is a specialized form of information retrieval. Given a collection of documents, a Question Answering system attempts to retrieve correct answers to questions posed in natural language.

QA in medical field is one of the more difficult compare to other field. Work in this field starts with difficulty of identifying the disease. When a user tells its question, the question mat contains not only the suffering disease but also the past disease and some symptoms, which may alone be a disease. E.g. fever can be a symptom and also a suffering disease. It is easy for human to understand the disease among all the information but difficult for a machine to identify.

E.g. *I am suffering from bipolar disorder since year 2000 ,at that time i got a maniac attack,after that i did not get any episodes for next 3-4 years then in year 2006 i got depression ,and in year 2007 again depression,then in year 2012 out of the blue moon i got maniac attack .Any suggestion.*

A human can easily from above information that patient is suffering from bipolar disorder. But, a machine needs to analyse and identify among bipolar disorder, maniac attack and depression.

Problem is not only identification of disease but also the length of question. A System performs better and provides efficient result, as the questions falls shorter and shorter. Length of questions increases; efficiency of results starts to decrease. In present time, many systems are available but they are restricted to some length of question, like 20-25 words. And sometimes, a user feels free to give their description in detail and feels inconvenience if they find any restriction. So, a system is required which should be independent of word length.

In recent time, internet users have increased drastically. Now a day, lots of normal users are accessing the web based health information system, which are unable to write formal English. They give the information in informal text and system does not identify the informal text. For medical domain, each word is important because based on these words only the system provides the result. If the system fails to identify the informal text and correct them, system does not analyze the correct question and fails to answer the question.

Even if we overcome these restrictions and does not make a system which uses the internet information but uses its own database information, we cannot have faith for long time. Medical information keeps on updating time-to-time, based on recent development. If we use a database, we need to update it continuously and needs to be in touch with all development that occurs, which is always not possible. But also with the use of internet, main concern is authentic information. For a disease, a website provides their information may differ from others, so it is also necessary to choose the best from all these. Also, we cannot rely on a single website for all disease information, because it is possible that they do not have information for all disease.

A websites along with the required information provides a number of irrelevant information which may be not required by the user. So the system should be able to identify and select only the specific ones. Selection of special kind of data and rejection of other data is also very hard to execute.

Usually, when a patient gets ill, it may take some time to meet the doctor. Meanwhile, the situation of the patients becomes adverse, which can be better if he/she gets the correct precautionary information. If the user gets prior knowledge before meeting to doctor about the medical test (which may be required), patient can save their precious time and in the first meet to doctor, treatment will get start.

Today, a user has become so clever that he gets all the information about the problem before going to the doctor, so can get the better treatment. Normally, physician gives the medicine to the patient but not the information. A patient knows that he gets only the medicine for the

Manuscript received 19 July, 2015

Premchand, B.Tech from B.I.T Sindri, Dhanbad and pursuing M.Tech from Shree Institute of science & technology, Bhopal, India
Yogesh Rai, Working with Shree Institute of science & technology, Bhopal, India

concerning problem. So he looks for the disease related information that helps to prevent the same problem again.

II. SYSTEM MODEL

In this report, we focus on natural language medical question in larger length. Our main concern is: when a long question is given, system should be able to identify the medical words correctly and should be able to provide the best health information based on question. In order to achieve this goal, we generally separate the question answeringprocedure into four steps: question identification, question analysis, web search, and information extraction (shown at Fig.1).

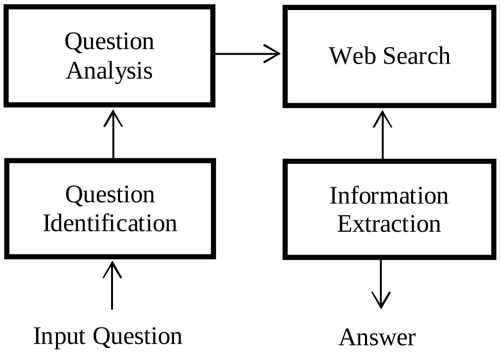

Fig 1: Generic structure of QA system

A. Question Identification:

This step mainly focuses on identification of question and rectification of wrongly spelt words. If the system is closed-domain based, it cannot be performed on all questions. If performed, system would fail to result any answer. But if results, the system will take much longer time and answer will be inaccurate. The system should be able to identify the nature of question and execute the question only if it belongs to its domain. Also, the system should be able to identify the wrongly spelt words (if any) and either correct those automatically or gives the option to user to correct manually.

B. Question Analysis

Once the question identification step is performed, the correct question is accepted by the system and further performed for analysis.This part is the soul of the system. This is so called because identification of medical terms and their classification is done, based on which final result depends. All the important information is extracted and they are classified according to their property. The main work here is correctly identification of disease and symptoms, which is the toughest job of the project.

C. Web Search

The whole question cannot be given to the website for result, because the website will extract some (important) words from the question based on its own strategy and will give the results based on these words (all the websites always give the result based on the important words from the question). So, in order to avoid this situation, query can be made based only on important words selected by the system. The query can be formed by two ways: from top most frequent and relevant words/phrases or from the template information.

Now the query is searched on the internet to get the answer of the question asked and the web results are considered as the webpage which contains the answer of the question. When search is performed on the internet, the search engine gives a numerous of webpages. Generally, all the resulted webpages does not possess the required information. Here main job is to access the webpage which contains the required information.

D. Information Extraction

In order to achieve more efficient result, the question query can be searched over more than one website and their results can be used in some manner to get more specific results. The searching of query in more than one website holds better result because sometimes a website fails to provide any medicine name for a disease and in that case other website result can be helpful.

In the selected webpages, there may a lot of contents other than the required information. The webpage may contain: introduction of specific terms, several non-relevant links, unrequired images, many advertisements, etc. So, to be more accurate about the answer, the retrieved information must be more specific and towards the question.

Same type of information can be available in more than one webpages. System should not present all same type of information as result but only unique data. So needs to select the better information among all, which can be represented as final.

III. LITERATURE REVIEW

Work of question answering system mainly focuses identification of text i.e. named entity recognition (NER) and information extraction. During our survey, we have come across a number of research papers and present an overview.

A. NER in the Biomedical Domain

NER (Named Entity Recognition) is a process that identifies atomic elements in text and classifies into predefined categories. NER is concerned with recognition of medical entities in text, such as recognition of gene/protein names in PubMed abstracts. We review NER in the biomedical domain based on three major features: entity types, features, and techniques.

Entity types: A named entity is a word or a phrase that refers to a particular kind of object. NER in the medical domain generally target medical substances, diseases, and medical functions. A number of NER tools have been developed for medical substances name recognition, such asABNER (Settles, 2005), POSBIOTM—NER (Song, Kim,Lee,&Yi, 2005), and NERBio (Tsai et al., 2006). Some standard test corpora such as the GENIA corpus (Team Genia, 2006) and the BioCreative corpora (Krallinger, 2006) also have been created to allow researchers to compare the performance of each NER technique for medical substances entity type.

Features: The quality of NER model depends on the power of the data representation; i.e. the features, which can be classified into three types: lexical, syntactic and semantic features.

The lexical feature type considers the word appearance such as numerical letter, uppercase letter, punctuations, and word morphology. Many named entities contain noticeable lexical patterns. E.g., many medical processes have common suffixes like -sis, -ism, and –ion (Collier, Nobata, & Tsujii, 2000; Tsai et al., 2006).

Syntactic features use structural properties of the text. Syntactic featuresare more flexible and richer, and therefore harder to model, than lexical. In particular, part-of-speech (POS) tags, a set of syntactic features, have been used in a number of NER studies (Nichalin Suakkaphong et al., 2011; Tsai et al.,2006; Zhou, Zhang, Su, Shen, & Tan, 2004).

Semantic features takes the meaning of word and the word belongs to certain predefined semantic classes. Semantic features are highly informative, but mostly suffer from poor coverage and feels lack of portability across different named entity types. It may require human effort in maintaining the word lists or dictionaries. (Hatzivassiloglou et al. (2001); Zhou et al. (2004).

Techniques: Techniques for NER can be classified into two categories: knowledge-based approaches, and machine learning.

Knowledge-based approaches are commonly adopted by linguistic researchers and they use of human knowledge in developing grammar rules and dictionaries. Such approaches may require very little or none training data, yet developing and maintaining a knowledge-rich system is typically time-consuming and expensive.

In spite of the availability of well-maintained knowledge sources, the performance of knowledge-based systems is not consistently good. For instance, BioMedLEE (Borlawsky et al., 2006) has an average precision and recall of approximately 94 and 85%, respectively, for disease names while MetaMap (Pratt&Yetisgen-Yildiz, 2003) has an average precision and recall of 27.7 and 52.8%, respectively (One third of the MetaMap test data is disease names.)

Unlike knowledge-based approaches, machine learning approaches aim at reducing human effort in maintaining rules and dictionaries. NER is typically viewed as a sequence-labeling problem. In one of the approach, breaks the sequence into a number of data points and solves problems independently, using techniques such as support k-nearest neighbor, maximum entropy, or naïve Bayes. This tries to make the best local decision at each data point by minimizing labeling error. This is a common technique that has been used in many domains (e.g., biomedical relation extraction: Li et al., 2008; biomedical literature classification: Lertnattee & Theeramunkong, 2004). A recent advance along this line is large-margin classifiers with structured outputs, which takes output variables in an interdependent fashion.

B. Information Extraction

Bisharah Libbus et al., 2002, performed a project to construct a tool for identifying and extracting biomedical information from texts. With this tool, they are aiming to provide as much information as possible to the molecular biologists. Generally such a tool exploits terms and relations identified automatically in text by both statistical and symbolic methods in addition to information supplied by National Library of Medicine indexes.

For Disease and Findings identification, they have used Metamap, which is a program that maps texts to concepts in the Unified Medical Language System (UMLS) Metathesaurus. To identify variables values, they have produced a syntactic parse structure which identifies noun structure. Based on this result, Metamap maps the parsed sentences to concepts in UMLS Metathesaurus.

Ronen Feldman et al., 2002, described a structure driven rule based strategy. This extracts the predefined semantic relationships using syntactic and semantic analysis of sentences. Including to this, also describes another information extraction approach using generic syntax based templates. To save the time it consumes to write patterns for all possible lexical combinations and semantic combinations for certain relationship; they have taken a verbRelation Template by considering a collection of MEDLINE abstracts where all the relevant entities (Gene, Disease, Tissue etc.) are available.

Latha.Ket al. gathered 1000 sample sets of biomedical documents from PUBMED, MEDLINE and NLM. Then these documents are processed in tokenization, data cleaning, stop word removal, stemming and identification of interesting terms. Paice/Husk algorithm is used for stemming to produce more mean modified hamming distance than other algorithms. At next, Support Vector Machine algorithm is used to analyze the processed documents. Then several data mining methods are applied for the information extraction.

IV. PROBLEM FORMULATION

It is observed that different semantic and syntactic analysis is used for many for identification of entities in the question answering system. Many question answering system are performing well in medical domain, but still there is a gap in the technology. These systems are well efficient in performing the single sentence or small questions, but still starts decreasing their efficiency drastically with the increase in question length or multi-sentence question.

V. PROPOSED METHODOLOGY

In this work, we aim to develop a system, made for end user that is able to generate the most specific information user is searching for. The system is capable of handling very long query (like 100 words) of the user. Apart from the disease description, we also provide the tests required, probable symptoms and home remedies.

English and medical data dictionary can be used to identify the wrongly spelt words (if any) using edit distance method and options can be given to user to correct the word. Identification and selection of question can be done using data sets. Two data sets can be created one for close domain and one for open domain. Both datasethave same special words with some value for every word but different in both dataset.

We can analyze the user's (very long) query and from the question information, generate the template. Template contains the important words, like disease name, symptoms, age, sex, disease duration. We can use this information in query formation and can used for web search. Disease and symptoms can be differentiated using some rule based strategies. As use of few words is better for efficient web search, disease name and symptoms can be used to get much close information to user's question. From the web, we retrieve the information. Like other system, we do not use the user's question only at the beginning but also use it for the final selection of information. The health information accessed from various web pages are compared again with the help of the information from the user's question, to get the most specific answer.

VI. CONCLUSION

In this survey, we presented the exploratory and comprehensive review and theoretical analysis of different question answering techniques.Basically biological and biomedical texts are unstructured. These texts need to be made relational and structured so that they can be computerized. If formal dictionary is used, this problem can be lowered drastically. With the problem of long questions, template information can be used in different ways to reduce the question size. This small question can then be used for web information extraction and can eventually upgrade the efficiency of the system.

REFERENCES

[1] Libbus, Thomas C. Rindflesch, "NLP-Based Information Extraction for Managing the Molecular Biology Literature", AMIA 2002 Annual Symposium Proceedings

[2] Ronen Feldman, Yizhar Regev, Michal Finkelstein Landau, Eyal Hurvitz, Boris Kogan, "Mining Biomedical Literature using Information Extraction", ClearForest Corp, USA, Israel, October 2002.

[3] Latha.K, Kalimuthu.S, Dr.Rajaram.R, "Information Extraction from Biomedical Literature using Text Mining Framework", International Journal of Imaging Science and Engineering, GA, USA, ISSN: 1934-9955, VOL.1, NO.1, January 2007.

[4] Purabi Kalita, Rashmi Choudhury, " Information Extraction for Biomedical and Biological Literature", International Journal of Computer Applications (IJCA). National Conference cum Workshop on Bioinformatics and Computational Biology, NCWBCB- 2014.

[5] Settles, B. (2005). ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics, 21(14), 3191–3192.

[6] Song,Y., Kim, E., Lee, G.G., &Yi, B. (2005). POSBIOTM—NER:A trainablebiomedical named-entity recognition system. Bioinformatics, 21(11),2794–2796.

[7] Tsai, R.T., Sung, C.L., Dai, H.J., Hung, H.C., Sung, T.Y., & Hsu, W.L. (2006). NERBio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition.BMC Bioinformatics, 7(Suppl. 5), S11.

[8] Team Genia. (2006). GENIA corpus—Genia Project Homepage. Retrieved from http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page= GENIA+corpus

[9] Krallinger, M. (2006). BioCreAtIvE homepage. Retrieved from http:// biocreative.sourceforge.net/index.html

[10] [10] Hatzivassiloglou, V., Duboué, P.A., & Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: A machine learning approach. Bioinformatics, 17(Suppl. 1), S97–S106.

[11] Borlawsky, T., Friedman, C., & Lussier, Y.A. (2006). Generating executable knowledge for evidence-based medicine using natural language and semantic processing. In Proceedings ofAmerican Medical Informatics Association (pp. 56–60). Bethesda, MD: AMIA.

[12] Pratt,W.,&Yetisgen-Yildiz, M. (2003).A study of biomedical concept identification:MetaMap vs. people. In Proceedings of the Annual AmericanMedical Informatics Association Symposium (pp. 529–533). Bethesda,MD: AMIA.

[13] Lertnattee, V., & Theeramunkong, T. (2004). Multidimensional text classification for drug information. IEEE Transactions on Information Technology in Biomedicine, 8(3), 306–312.

Premchand has completed his B.Tech from B.I.T Sindri, Dhanbad and pursuing M.Tech from Shree Institute of science & technology, Bhopal, India
Pbertam2003@yahoo.com

Yogesh Rai has completed his B.E from LNCT, Bhopal and M.Tech from LNCT, Bhopal. Currently he is working with Shree Institute of science & technology, Bhopal, Ind